SVKM's
D. J. Sanghvi College of Engineering

Program: B.Tech in Computer Science and Engineering (Data Science) **Academic Year: 2022** **Duration: 3 hours**
Date: 25.01.2023
Time: 09:00 am to 12:00 pm
Subject: Foundations of Data Analysis (Semester III) **Marks: 75**

Instructions: Candidates should read carefully the instructions printed on the question paper and on the cover page of the Answer Book, which is provided for their use.

(1) This question paper contains_**03**_pages.
(2) All Questions Are Compulsory.
(3) All questions carry equal marks.
(4) Answer to each new question is to be started on a fresh page.
(5) Figures in the brackets on the right indicate full marks.
(6) Assume suitable data wherever required, but justify it.
(7) Draw the neat labelled diagrams, wherever necessary.

Question No.		**Max. Marks**
Q1 (a)	i. Explain any three type of probability sampling with suitable example of each. **OR** ii. Identify the suitable type of sampling technique from the below scenarios given also given proper justification: A. The researcher assigns every member in the company database a number. Instead of randomly generating numbers, a random starting point (say 5) is selected. From that number onwards, the researcher selects every, say, 10th person on the list (5, 15, 25, and so on) until the sample is obtained. B. The researcher stands outside a company and asks the employees coming in to answer questions or complete a survey. C. A company has over a hundred offices in ten cities across the world which has roughly the same number of employees in similar job roles. The researcher randomly selects 2 to 3 offices and uses them as the sample. D. If a company has 500 male employees and 100 female employees, the researcher wants to ensure that the sample reflects the gender as well. So the population is divided into two subgroups based on gender. E. The researcher wants to know about the experiences of disabled employees at a company. So the sample is purposefully selected from this population.	[05] [05]
Q1 (b)	i. Draw a star schema for the Education System data warehouse using the schema given below: (Assume suitable data wherever necessary) Dimensions: Time, Student, Course, Accounts, Department, Faculty Facts: No. of enrollments, no. of courses, no. of published papers, no. of rejected papers, course fee. ii. Explain various modes of applying data in data warehouse using suitable diagram.	[06] [04]
Q2 (a)	Explain the various types of data sets and justify with suitable example of any 2 data set types.	[05]

Q2 (b)

i. Yashree is a good student, but at times she doesn't get enough sleep. She hypothesizes that when she gets more sleep she does better on tests. To test her hypothesis, she tracked how she did on a number of tests, based on how many hours of sleep she got on the night previous. Find the value of the Pearson's correlation coefficient of between the two variables. [10]

Hours of Sleep	Test Score
8	81
8	80
6	75
5	65
7	91
6	80

OR

ii. Explain the various univariate methods in feature engineering, also calculate the Signal to Noise Ratio for the data: 1, 5, 6, 8, 10. [10]

Q3 (a)

i. Explain the types of anomalies in data preprocessing with an example of each. [08]

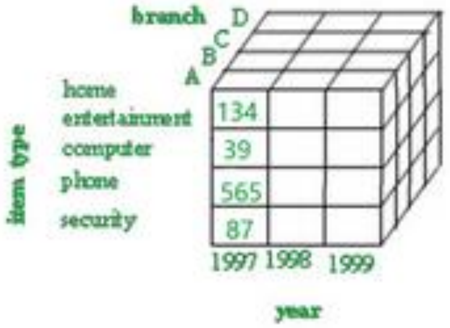
OR

ii. Consider the data for price (in euros): 8, 30, 3, 13, 22, 26, 22, 26, 28, 7, 37, 22 apply the binning by using mean, median and boundaries technique where data in each bin is four. [08]

Q3 (b)

i. What is the need of storing data in a data cube in data warehouse? [03]

ii. Perform OLAP operation for the given data cube: [04]



A. Which OLAP operation you will use to analyze data for the year 1997. Draw new OLAP cube for the same.

B. Which OLAP operation you will use to analayse data for item_type = "Entertainment" and "Phone" and year = 1997.

Q4 (a)

i. Choose the appropriate answer from the following and give proper justification for the same. [05]

1) What type of join is used in blending?
   a) Right Join b) Left Join c) Full Join d) Inner Join
2) Which graph in visualization depicts the data in a color-coding technique for the different values of data?
   a) Line Graph b) Heat Map c) Scatter Plot d) Pie Chart
3) Which of the following is used to show 2 measures in a single graph?
   a) Label b) Detail c) Dual Axis d) Color
4) Which of the following is rightly used to show the distribution of continuous information over a certain period of time?
   a) Bar Graph b) Line Chart c) Pie Chart d) Histogram
5) Which of the following is rightly composed of multiple bars stacked vertically one on another?
   a) Line Graph b) Pie Chart c) Stacked Bar Graph d) Bar Graph

OR

ii. Select the appropriate attribute with the type of data(s): [05]

Attributes	Type of Data
A. Gender (M,F)	1. Nominal
B. No. of students in a class	2. Ordinal
C. Rank (1, 2, 3)	3. Discrete
D. Height	4. Continuous
E. The no. of workers in a department	
F. Hair Color (Blonde, Brown)	
G. The sq.ft. of a house	
H. The speed of car	
I. The no. of home runs in basketball	
J. Letter (A, B, C)	

Q4 (b) i. Consider the following data points and find if A is an outlier point using Local outlier factor method with K=3. [10]
A= (7,5), B= (1,5), C= (1,4), D= (4,1), E= (3,1), F= (3,0) and G= (4,0)
[Hint: use Manhattan distance for distance calculation]

OR

ii. A survey was given to a random sample of 20 sophomore college students. They were asked, “how many textbooks do you own?” Their responses, were: 0, 0, 2, 5, 8, 8, 8, 9, 9, 10, 10, 10, 11, 12, 12, 12, 14, 15, 20, and 25. Compute the IQR and also find out the outliers in the data given [10]

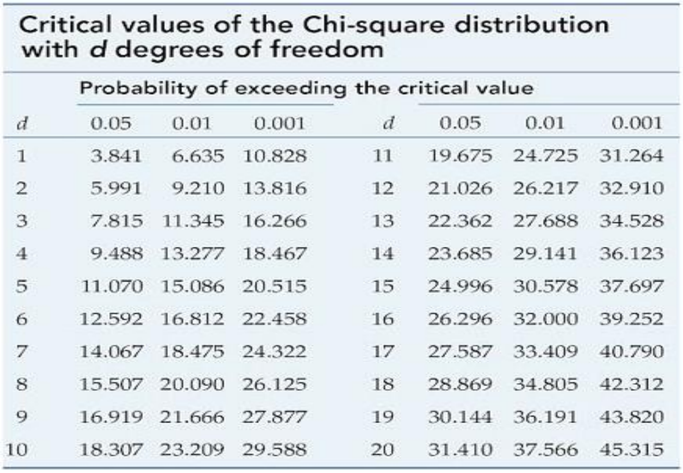
Q5 (a) Explain the data warehouse components with suitable diagram. [07]

Q5 (b) i. Explain various multivariate methods used for feature selection. Find principal components for the given dataset: (2,1), (3,5), (4,3), (5,6) (6,7) and (7,8). [08]

OR [08]

ii. Calculate the Chi-square value for the following data of incidences of water-borne diseases in three tropical regions.

	India	Equador	South America
Typhoid	31	14	45
Cholera	2	5	53
Diarrhoea	53	45	2



Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315